

Cell Type–Specific Biomarkers of Systemic Sclerosis Disease Severity Capture Cell-Intrinsic and Cell-Extrinsic Circuits

Jacob S. Berkowitz,¹ Tracy Tabib,² Hanxi Xiao,¹ Gabrielle M. Sadej,² Dinesh Khanna,³  Patrizia Fuschiotti,² Robert A. Lafyatis,² and Jishnu Das¹ 

Objective. Systemic sclerosis (SSc) is a multifactorial autoimmune fibrotic disorder involving complex rewiring of cell-intrinsic and cell-extrinsic signaling coexpression networks involving a range of cell types. However, the rewired circuits as well as corresponding cell–cell interactions remain poorly understood. To address this, we used a predictive machine learning framework to analyze single-cell RNA-sequencing data from 24 SSc patients across the severity spectrum as quantified by the modified Rodnan skin score (MRSS).

Methods. We used a least absolute shrinkage and selection operator (LASSO)–based predictive machine learning approach on the single-cell RNA-sequencing data set to identify predictive biomarkers of SSc severity, both across and within cell types. The use of L1 regularization helps prevent overfitting on high-dimensional data. Correlation network analyses were coupled to the LASSO model to identify cell-intrinsic and cell-extrinsic co-correlates of the identified biomarkers of SSc severity.

Results. We found that the uncovered cell type–specific predictive biomarkers of MRSS included previously implicated genes in fibroblast and myeloid cell subsets (e.g., SFPR2+ fibroblasts and monocytes), as well as novel gene biomarkers of MRSS, especially in keratinocytes. Correlation network analyses revealed novel cross-talk between immune pathways and implicated keratinocytes in addition to fibroblast and myeloid cells as key cell types involved in SSc pathogenesis. We then validated the uncovered association of key gene expression and protein markers in keratinocytes, KRT6A and S100A8, with SSc skin disease severity.

Conclusion. Our global systems analyses reveal previously uncharacterized cell-intrinsic and cell-extrinsic signaling coexpression networks underlying SSc severity that involve keratinocytes, myeloid cells, and fibroblasts.

INTRODUCTION

Fibrotic skin disease is a sine qua non of diffuse cutaneous systemic sclerosis (dcSSc), the disease subset more commonly associated with most severe complications, including scleroderma, renal crisis, joint contractures, and interstitial lung disease, as well as disease-related mortality. Disease markers correlating with skin disease severity, as assessed by the modified Rodnan skin score (MRSS), have provided insight into disease pathogenesis and biomarkers for clinical trials. In previous studies examining gene expression in RNA extracted from entire biopsies (bulk RNA), we have shown that markers of fibroblasts, myofibroblasts, and macrophages correlate with the MRSS (1–3). These studies

suggested that macrophage and fibroblast cell populations are particularly important in promoting disease. Multiple other investigators have analyzed bulk skin gene expression for defining patient “subsets” and for investigating disease pathogenesis (4,5). Although algorithms have been developed to deconvolute bulk gene expression, this approach cannot clearly identify the heterogeneous fibroblast, vascular, epithelial, glandular, and inflammatory cell populations in the skin (6,7). In addition, bulk gene expression cannot tell if a gene is increased because a stable cell number is expressing higher levels or the cell population expressing a gene has expanded in number.

Single-cell RNA sequencing provides a more granular view of gene expression in skin. Clustering of cells enables understanding

Supported by the University of Pittsburgh Center for Research Computing through the high-performance computing resources provided. The work of Drs. Lafyatis and Das was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH, (grant P50-AR-060780).

¹Jacob S. Berkowitz, BS, Hanxi Xiao, MS, Jishnu Das, PhD: Center for Systems Immunology, Departments of Immunology and Computational & Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania; ²Tracy Tabib, MS, Gabrielle M. Sadej, BS, Patrizia Fuschiotti, PhD, Robert A. Lafyatis, MD: Division of Rheumatology and Clinical Immunology, Department of

Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania; ³Dinesh Khanna, MBBS, MSc: Division of Rheumatology, Department of Medicine, University of Michigan, Ann Arbor.

Author disclosures and a graphical abstract are available online at <https://onlinelibrary.wiley.com/doi/10.1002/art.42536>.

Address correspondence via email to Robert A. Lafyatis, MD, at lafyatis@pitt.edu, or to Jishnu Das, PhD, at jishnu@pitt.edu.

Submitted for publication September 16, 2022; accepted in revised form April 13, 2023.

of average gene expression in various cell populations. By comparing SSc to control cells in the same population, we and others have elucidated corresponding transcriptomic changes within cell types of interest (8–10). In this study, we sought to use single-cell RNA sequencing coupled with multivariate machine learning approaches to delve deeper into cell type-specific changes that characterize patients across the disease severity spectrum. These analyses provide a deeper understanding of cell type-specific biomarkers of disease severity, beyond correlates that stratify disease versus healthy individuals. These comparisons also focus on granular cell-intrinsic as well as cell-extrinsic factors and associated interactions linked to disease severity.

We conceptually extended a predictive machine learning approach that we had previously used to identify correlates of natural and vaccine-induced immunity in a wide range of infectious disease contexts including HIV (11–14), malaria (15,16), and tuberculosis (17). Our approach used a 2-step feature selection strategy that takes into account the high dimensionality of the data set and helps avoid overfitting via the use of stringent L1 regularization. Our goal was to determine a minimal number of cell type-specific biomarkers that best predict SSc severity as quantified by the MRSS. While our cohort included healthy controls, they were not used in the identification of biomarkers of SSc severity. This is because we were well powered to identify biomarkers of SSc severity, but not adequately powered with this cohort to identify differences between healthy individuals and SSc. Furthermore, correlates that can discriminate between SSc and healthy individuals are already well characterized. Our study focuses on the more difficult and underexplored problem of identifying robust multivariate cell type-specific signatures of SSc severity.

PATIENTS AND METHODS

Single-cell RNA-sequencing analyses and data preprocessing. We obtained single-cell RNA-sequencing data for 24 patients with dcSSc. These patients have been previously described (9,10). Single-cell RNA-sequencing data from 24 SSc patients, all obtained using the same 10X Genomics V2 chemistry, were processed using a standard analytic pipeline comprising alignment via Cell Ranger and postprocessing (dimensionality reduction and clustering) using Seurat version 3.2.2. We used a systematic set of filters to remove unhealthy cells, dropouts, and doublets. For the entire data set, we only retained cells that expressed ≥ 200 different genes and had $<35\%$ mitochondrial content. We also filtered to retain cells with $<5\%$ hemoglobin content. These broad filters were applied to all cell types. We also applied the following additional filters to the primary cell types, including myeloid cells, fibroblasts, and keratinocytes. To remove cells with ambient keratin RNA, we filtered the cells so that none had a percentage of "KRT" gene expression $>0.5\%$. We also

further filtered the percentage hemoglobin (percent.hemo) to only include cells that had $<1\%$ hemoglobin content. The percentage mitochondrial filter was also more stringently set to only include cells with $<10\%$ mitochondrial content. We also further filtered so that a cell had to have $\geq 1,000$ RNA molecules, but $<40,000$ RNA counts to be retained.

Data analysis was performed using R version 4.00. The clusters were manually assigned based on marker genes. Seurat was used for identification and visualization of cell populations and for obtaining normalized gene expression, additionally using the "AverageExpression" function to calculate the average expression levels of each gene within an individual cluster per patient. Clusters with an across-patient average of <20 cells were excluded from further analysis, as well as the neural cluster.

Least absolute shrinkage and selection operator (LASSO) model utilizing gene expression levels across all cell types. For each of the 17 clusters, we focused on the top 20 high-variance genes (variance calculated across patient samples). The variance was calculated in an unsupervised fashion, i.e., without looking at labels (MRSS) for these samples. Thus, we ended up with 340 features for each of the 24 patients. The 20 high-variance genes are unique within their given cell cluster. However, if a specific gene was selected for 1 cluster, it was still eligible for selection in other clusters, as some genes could be important in >1 cluster. We then used a LASSO-based model with this reduced set of features to identify a minimal set of predictive biomarkers of outcome (MRSS). The model consisted of LASSO-based feature selection followed by regression using the down-selected features using a random forest model. Thus, the LASSO regression coefficients were simply used in an off/on fashion, i.e., features with non-zero coefficients in the LASSO model were selected and passed to the random forest model. A random forest using the down-selected features was used for the actual prediction of MRSS. The use of L1 regularization helps avoid overfitting on a high-dimensional data set—as even with the reduced set of features, the number of features far exceeds the number of samples—and drives sparsity. The objective function used for model tuning was the Spearman's correlation coefficient between true and predicted MRSS. The regularization parameter (lambda) was tuned using nested cross-validation. We then evaluated the performance of the model in a leave-1-out cross-validation framework with permutation testing as previously described (11,14–16). We also reported the "final set" of biomarkers using the entire data set (using the same lambda that we identified from nested cross-validation). These genes also correspond to the most stable/frequent genes across folds and replicates of cross-validation.

We also built an alternate model where we went through the same modeling and validation framework with cell type-specific differentially expressed genes (DEGs) rather than high-variance genes for the different cell types.

LASSO model utilizing gene expression levels in specific cell types. We also built cell type-specific models. For both fibroblasts and keratinocytes, we focused on the top 300 high-variance genes (variance calculated across patient samples). Thus, we ended up with 300 features for each of the 24 patients in each cell type. Independent models were built for each cell type. The model consisted of LASSO-based feature selection followed by regression using the down-selected features using a random forest model. Model performance was evaluated in a leave-1-out cross-validation framework with permutation testing as previously described (11,14–16).

Correlation networks around LASSO-selected biomarkers. The LASSO models described identify a minimal subset of predictive biomarkers. To reconstruct the entire set of cell type-specific transcripts correlated with outcome, we used a correlation network approach as previously described (11,14–16). Briefly, we calculated all pairwise correlations involving the LASSO-selected biomarkers for each model and visualized only the significant ones (significance defined as $|r| > 0.5$ and false discovery rate < 0.05) using a correlation network.

Immunohistochemistry. Multicolor staining was performed on formalin-fixed, paraffin-embedded skin samples using a tyramide signal amplification kit (Thermo Fisher). The antibodies employed in these experiments included S100A8 (product no. ab92331; Abcam) and KRT6A (product no. ab52120; Abcam). Confocal images were captured on an Olympus Fluoview 1000 confocal microscope using an oil-immersion 100 $\times$ objective.

Ethics approval and data availability. Ethics approval was not applicable to our study. We obtained single-cell RNA-sequencing data for 24 dcSSc patients. These patients have been previously described (9,10). Patients and/or the public were not involved in the design, conduct, reporting, or dissemination of this research.

Detailed code, associated data sets, and documentation are available at the authors' Github page at https://github.com/jishnu-lab/SSc_biomarker.

RESULTS

Identification of cell types in SSc patients. We sought to identify a minimal number of cell type-specific biomarkers that best predict SSc severity as quantified by the MRSS for the severity of skin disease in patients with dcSSc using a robust single-cell RNA-sequencing data set (Figure 1). Patient demographics and clinical features are described in Supplementary Table 1, available on the *Arthritis & Rheumatology* website at <https://onlinelibrary.wiley.com/doi/10.1002/art.42536>.

We identified 35 unique clusters, consistent with prior expectation (Figure 2A and Supplementary Figures 1 and 5, <https://onlinelibrary.wiley.com/doi/10.1002/art.42536>). For downstream analyses, we filtered out clusters containing < 20 cells in each of the 24 patients. We examined the expression distributions of the marker genes for each cluster, i.e., the cell type-specific DEGs in the remaining 17 clusters (Figure 1). Consistent with expectation, most of the cell type-specific DEGs were uniquely expressed in that cluster alone, with a few of these genes being expressed in multiple clusters (Figure 2B and Supplementary Figures 1 and 4). We examined changes in gene expression in each population of cells as well as markers of each cell population more likely to reflect change in cell numbers associated with disease severity.

Cell-extrinsic predictive biomarkers of SSc severity.

We first sought to identify cell-extrinsic biomarkers of SSc severity (as quantified by MRSS) using a predictive machine learning approach. We used 20 high-variance genes for each of the clusters as inputs to our model (see methods). We hypothesized that genes that show variability across patients might serve as good biomarkers of SSc severity. However, the high-variance genes were chosen in an unsupervised fashion to avoid any bias or signal leakage (see methods). Analogous to techniques we have used in the past, we then used LASSO-based feature selection followed by classification using the down-selected features (11–16). The use of L1 regularization helps avoid overfitting on high-dimensional data. Our model was able to accurately predict MRSS; it performed significantly better than a model built on shuffled data (i.e., permutation testing) in a cross-validation framework (Figure 3A).

We applied cross-validation as a rigorous way to evaluate the predictive performance of the model with data left out. The model identified a minimal set of 8 cell type-specific predictive biomarkers that are sufficient to accurately predict MRSS (Figure 3B). The biomarkers included correlates of both low and high disease severity, and included gene biomarkers from fibroblasts, myeloid cells, pericytes, and keratinocytes (Figure 3B). Interestingly, several of the LASSO-selected biomarkers were expressed in multiple cell types (Figure 3C). This suggests that although the model selected these genes in a specific cell type, in some cases their expression levels were correlated across multiple cell types, and the biomarker serves as a surrogate for the entire correlation block. Thus, these LASSO-selected biomarkers encapsulate both cell-intrinsic and cell-extrinsic correlates of SSc severity. Finally, we visualized the power of the down-selected biomarkers in stratifying patients by disease severity (Figure 3G) using partial least-squares discriminant analyses. We observed that the down-selected biomarkers provided clear stratification by disease severity (Figure 3G). The earlier cross-validation-based analyses demonstrate the predictive power of the identified features with data left out. This analysis helps visualize a signature from the whole data

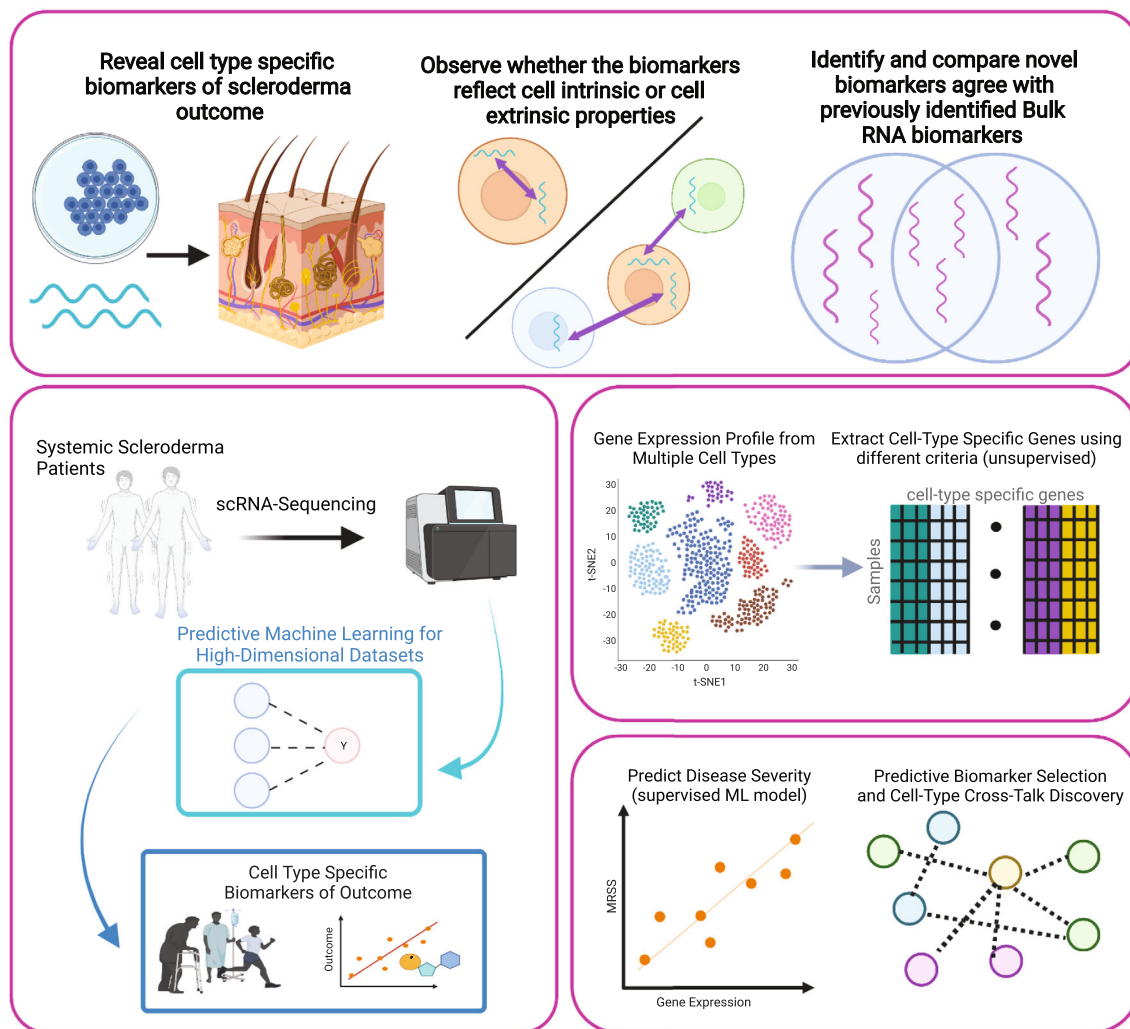


Figure 1. Schematic representation of the current study using a least absolute shrinkage and selection operator (LASSO)-based predictive machine learning (ML) approach on a single-cell RNA (scRNA)-sequencing data set to identify predictive biomarkers of systemic sclerosis severity, both across and within cell types.

set that can be utilized in downstream applications and that could be extrapolated to other cohorts.

Next, we sought to compare our predictive model to a model built on cell type-specific DEGs rather than high-variance genes. Analogous to our earlier feature construction approach, we used the top 20 cell type-specific DEGs as inputs to the LASSO model. The model built on cell type-specific DEGs remained significantly predictive in a cross-validation framework with permutation testing (Figure 3D). As in the earlier analysis, the biomarkers included genes from multiple cell types, including fibroblasts, myeloid cells, pericytes, and keratinocytes (Figure 3E). However, by construction, most of these biomarkers were specific to cell type (Figure 3F). A partial least-squares discriminant analysis using only these genes also stratified patients by disease severity (Figure 3H). However, the performance of this model was lower than the earlier model built using high-variance genes (Figure 3A). Our results suggest that models that incorporate

gene biomarkers variably expressed within a cell type and/or across multiple cell types are more predictive than models built on cell-specific gene biomarkers. Thus, particular genes in specific cells that manifest high variability in expression across patients, rather than genes that define cell types, more strongly influence clinical manifestations of skin disease.

To gain more holistic insights into the spectrum of transcripts tracking with SSc severity, we constructed correlation networks with stringent effect size and false discovery rate control, as previously described (13,15,16), around the LASSO-selected biomarkers (Figure 3I). Since LASSO identifies only a minimal set of predictive markers, the construction of correlation networks around the LASSO-selected biomarkers enabled us to identify co-correlates that are also associated with SSc severity. The correlation networks included biomarkers from across cell types, including T cells and endothelial cells, which were not included in the initial LASSO selection (Figure 3I). Intriguingly, in

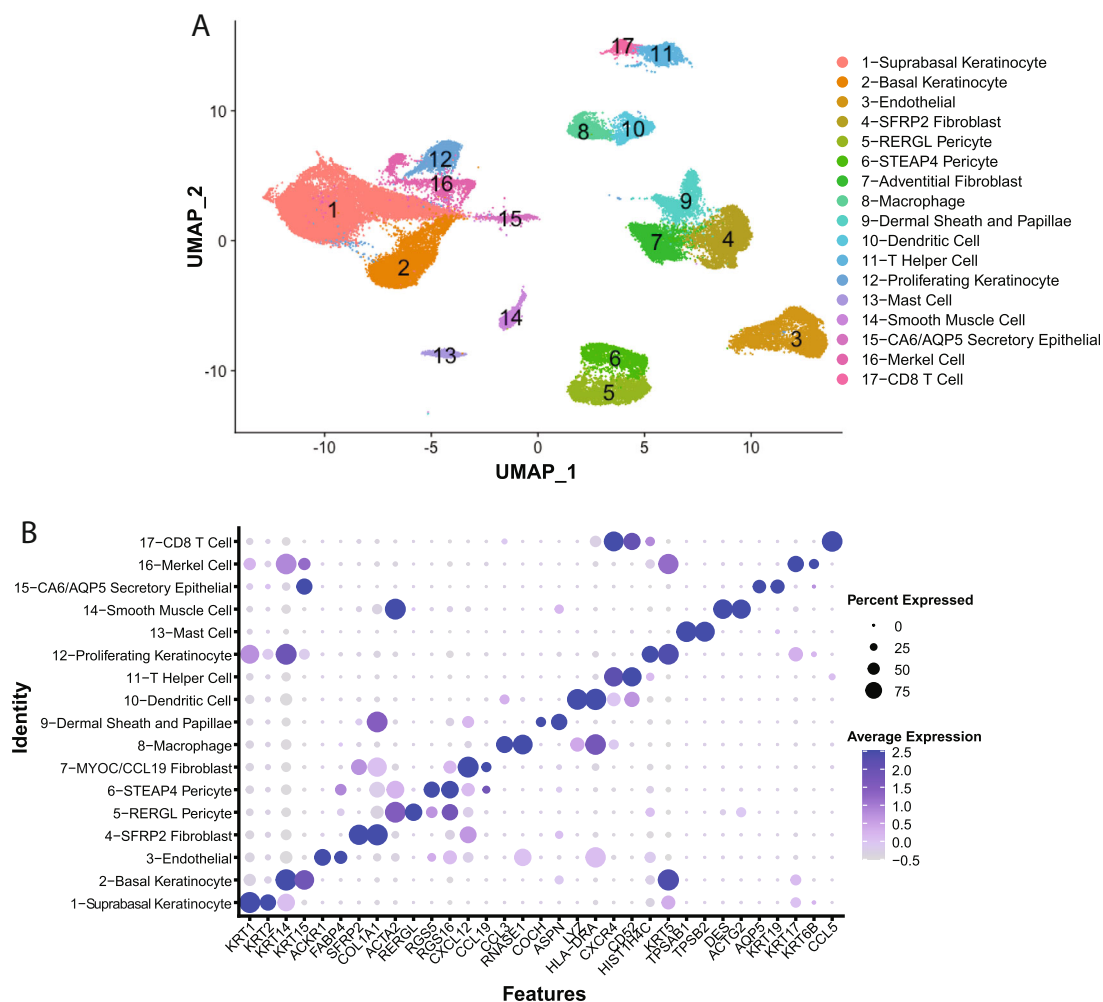


Figure 2. Cell types identified using single-cell RNA-sequencing data from the 24 systemic sclerosis patients. **A**, Uniform manifold approximation and projection (UMAP) analysis illustrating cell types identified using single-cell RNA-sequencing data from the 24 systemic sclerosis patients. **B**, Dot plot in which the size of the dot represents the percentage of cells within a cell type, while the color represents the average expression level. The top 2 cell type-specific differentially expressed genes are included in the plot.

addition to the same gene being, at times, correlated across cell types, genes across different cell types were also correlated (Figure 3I). These results provide insights into a broad range of cell type-specific transcripts associated with SSc severity. They also point to potential interactions between cells. For example, SAA1 expression by SFRP2-expressing fibroblasts correlated directly with pericyte expression of CXCL3, MYOC/CCL19-expressing fibroblast expression of CCL19, and proliferating keratinocyte expression of KRT6A. These and other co-correlates between cell populations suggest either a common factor is stimulating different gene expression in the various cell types, or that a disease-associated change in 1 cell type is acting upon and inducing the observed altered gene expression by the other cell type(s).

Cell-intrinsic predictive biomarkers of SSc severity.

While our earlier model provides insights into biomarkers of SSc severity spanning several cell types, we hypothesized that a

model that delves deeper into relevant cell types may perform better than a model in which the input features span multiple cell types and focus on the most variable genes across cell types. To test this hypothesis, we built 3 separate models using genes from only keratinocytes, fibroblasts, or myeloid cells. We chose these 3 cell types based on prior evidence implicating fibroblasts and myeloid cells in SSc and our results (shown in Figure 3) suggesting that several genes expressed in keratinocytes track with SSc severity.

Interestingly, each of the models built using genes from keratinocytes and fibroblasts were significantly predictive of SSc severity (Figures 4A and 4C). However, the model built on myeloid cells was not significantly predictive of SSc severity (Supplementary Figure 4, <https://onlinelibrary.wiley.com/doi/10.1002/art.42536>), suggesting that keratinocyte- and fibroblast-specific genes are better biomarkers of SSc severity than myeloid-specific genes.

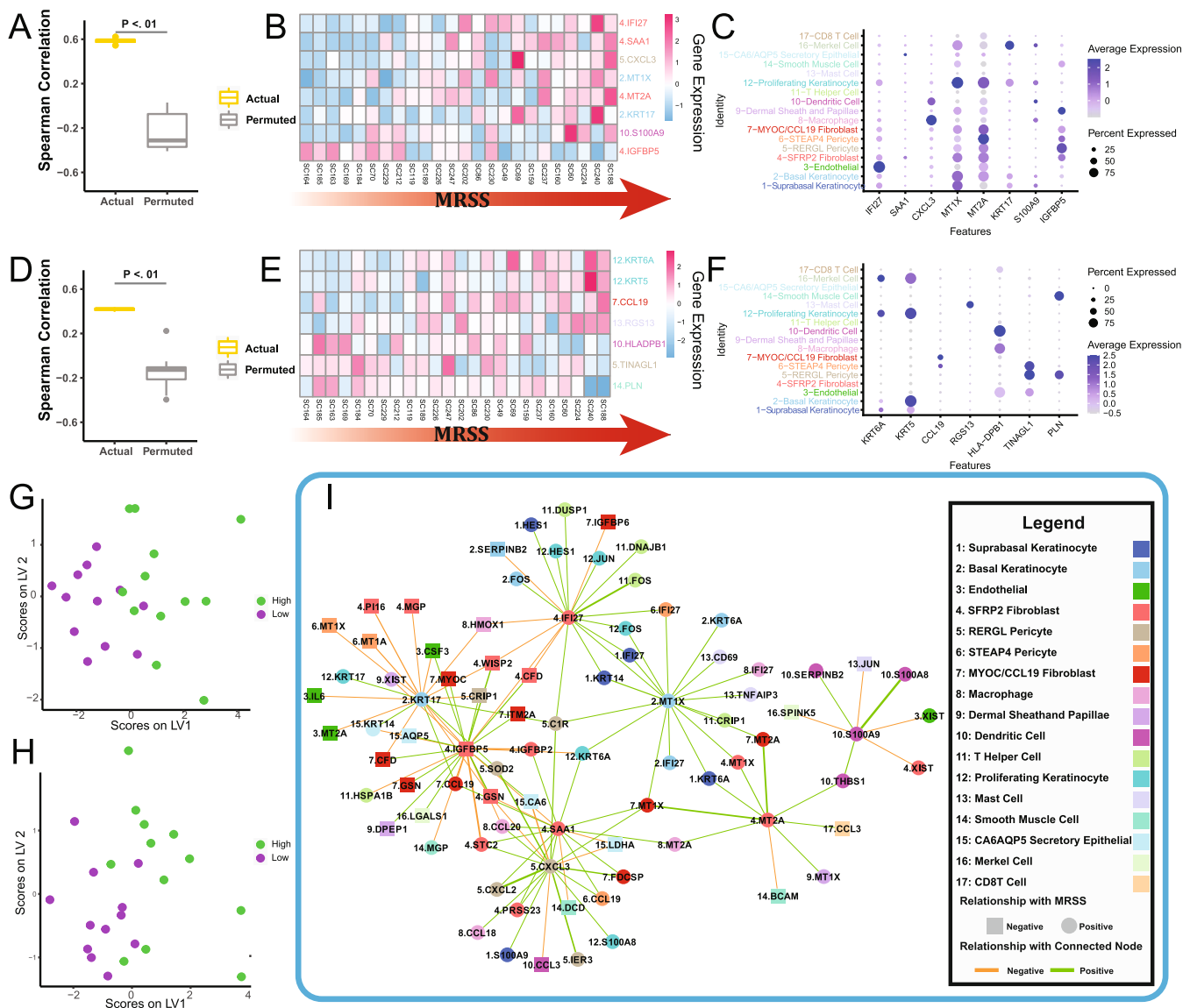


Figure 3. Uncovering robust biomarkers of systemic sclerosis (SSc) severity using pan cell-type LASSO models. **A**, Pan cell-type LASSO model for SSc severity biomarkers constructed with high-variance genes. Performance was measured using leave-1-out cross-validation. Spearman's correlation coefficients between predicted skin scores and true skin scores plotted. Permutation testing was used to assess model significance. **B–C**, Heatmap of expression levels of genes significantly predictive of modified Rodnan skin score (MRSS) among high-variance genes across samples (**B**) and dot plot of their expression levels across cell types (**C**). **D**, Pan cell-type LASSO model for SSc severity biomarkers constructed with cellular identity markers. Performance was measured using leave-1-out cross-validation. Spearman's correlation coefficients between predicted skin scores and true skin scores plotted. Permutation testing was used to assess model significance. **E–F**, Heatmap of expression levels of genes significantly predictive of MRSS among cellular identity markers across samples (**E**) and dot plot of their expression levels across cell types (**F**). **G–H**, Partial least-squares discriminant analysis constructed using the biomarkers of SSc severity among high-variance genes across cell types (**G**) and among cellular identity markers across cell types (**H**). High and low values were defined as greater than or less than the median, respectively. **I**, Correlation network illustrating significant co-correlates of the biomarkers of SSc severity identified in the model in **A**. In box plots, boxes show the median and interquartile range (IQR). Whiskers indicate 1.5 times the upper and lower IQRs. Symbols represent outliers. See Figure 1 for other definitions.

Notably, the keratinocyte-specific model used only 3 biomarkers (Figure 4B). These 3 markers, KRT5, KRT6A, and APOE, are part of a tight correlation network (Figure 4E and Supplementary Figure 3) that includes other genes and whose expression levels track with SSc severity and are involved in a range of functional processes. ZFP36

encodes tristetraprolin, an AU-rich element RNA binding protein that is up-regulated in wounded skin and controls inflammatory cytokines as well as neoplastic processes in keratinocytes (18,19). S100A2 regulates keratinocyte differentiation through its interaction with p53 (20), and expression of the alarmin S100A8 is associated with increased

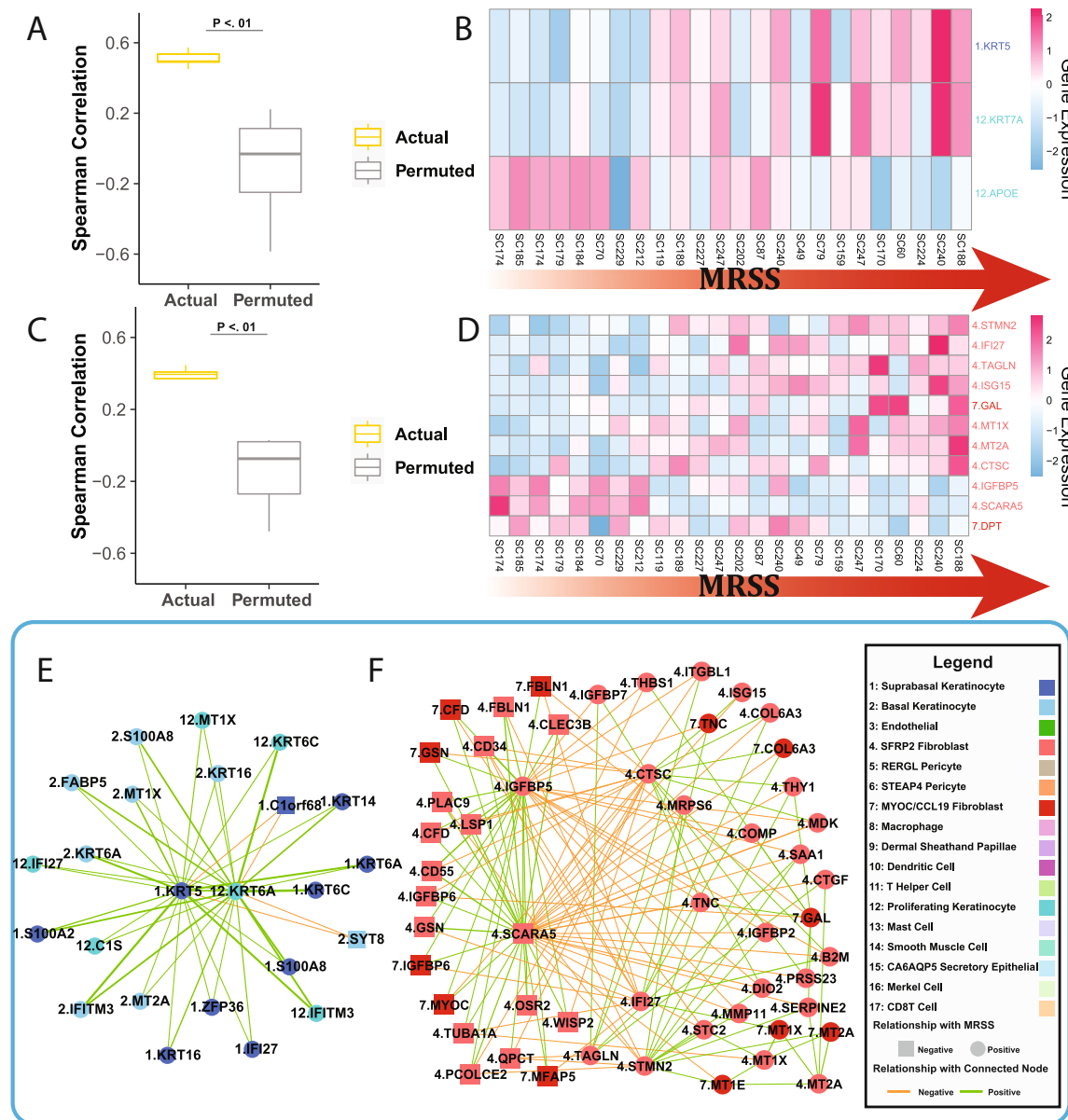


Figure 4. Uncovering robust biomarkers of systemic sclerosis (SSc) severity using cell type-specific LASSO models. **A**, LASSO model for SSc severity biomarkers specific to keratinocytes. Performance was measured using leave-1-out cross-validation. Spearman's correlation coefficients between predicted skin scores and true skin scores plotted. Permutation testing was used to assess model significance. **B**, Expression levels of genes significantly predictive of modified Rodnan skin score (MRSS) corresponding to the model in **A** across samples. **C**, LASSO model for SSc severity biomarkers specific to fibroblasts. Performance was measured using leave-1-out cross-validation. Spearman's correlation coefficients between predicted skin scores and true skin scores plotted. Permutation testing was used to assess model significance. **D**, Expression levels of genes significantly predictive of MRSS corresponding to the model in **C** across samples. **E–F**, Correlation network (restricted to degree ≥ 2) illustrating significant co-correlates of the biomarkers of SSc severity identified in the model in **A** (**E**) and biomarkers of SSc severity identified in the model in **C** (**F**). In box plots, boxes show the median and interquartile range (IQR). Whiskers indicate 1.5 times the upper and lower IQRs. See Figure 1 for other definitions.

keratinocyte proliferation and epithelial-mesenchymal transition (21).

The fibroblast-specific model included 11 biomarkers (Figure 4D), which were part of a much larger correlation network (Figure 4F and Supplementary Figure 3, <https://onlinelibrary.wiley.com/doi/10.1002/art.42536>). Many of the co-correlated network genes have been previously noted to correlate positively with the

MRSS, including CTGF, TNC, COL6A3, THBS1, COMP, THY1, PRSS23, and others (1–3,9,22,23). Coregulated expression of several of these genes was seen in both fibroblast subpopulations, such as TNC and COL6A3. As also seen in the mixed-cell model, several metallothionein genes were part of the correlation network. Notably, MT1X and MT2A were co-correlated in both fibroblast subpopulations. Metallothionein genes regulate metal

homeostasis, moderate oxidative stress, and block apoptosis (24,25). They are transcriptionally regulated by metal ions, stress, and hypoxia (26). A large number of other genes co-correlated negatively with the MRSS, including CD34, IGFBP5, IGFBP6, CD55, CFD, FBLN1, and others. Loss of CD34 has long been recognized as a marker of skin disease in SSc (27) and has more recently been associated with a transition to CD90/THY1-staining cells, and the ratio of this expression has been shown to correlate with the MRSS (22).

Surprisingly, the performance of the keratinocyte-specific model was better than the fibroblast-specific model. Our results suggest that while a larger number of genes in fibroblasts track with SSc severity, they individually, as well as collectively, associated less strongly with MRSS compared to a smaller number of keratinocyte-specific biomarkers (Supplementary Figure 2, <https://onlinelibrary.wiley.com/doi/10.1002/art.42536>). Intriguingly, although the above models are predictive of SSc severity, they do not perform as well as the overall model (Figures 3A, 3I, 4A, 4C). Our results suggest that while

individual cell type-specific biomarkers are informative of SSc severity, a pan cell-type analysis can identify more robust biomarkers. These findings reaffirm the need to perform pan cell-type analyses as both cell-extrinsic and cell-intrinsic biomarkers are associated with SSc severity.

Validation of the identified biomarkers of SSc severity. Next, we sought to compare the identified biomarkers from our single-cell RNA-sequencing data set with previously identified biomarkers from bulk RNA-sequencing data sets. This design achieves a 2-tier validation. First, this approach helps compare biomarkers identified from a single cohort in this study to correlates identified from a wide range of studies conducted by multiple groups. Second, it also evaluates whether the identified biomarkers are robust to the choice of platform used to identify them (i.e., bulk versus single-cell RNA sequencing).

We first identified a list of biomarkers identified from a spectrum of previous studies that had used bulk RNA sequencing to identify correlates of SSc severity (1–3,8,22,23,28). An early

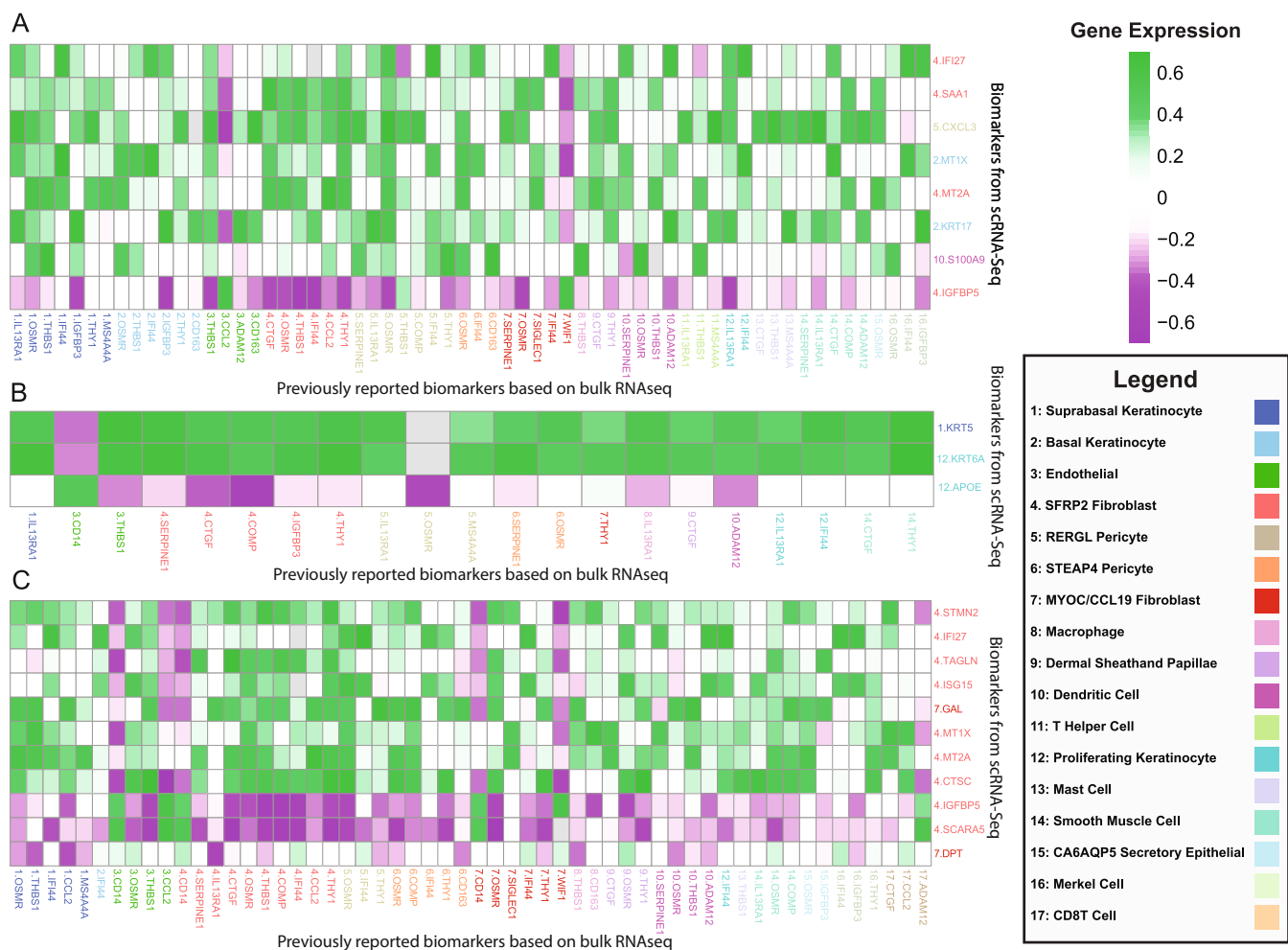


Figure 5. Comparison of cell type-specific biomarkers identified in this study to previously identified biomarkers from bulk RNA sequencing. **A–C,** Heatmaps illustrating correlations between the cell type-specific biomarkers identified in the pan cell-type model (**A**), in the keratinocyte-only model (**B**), or in the fibroblast-only model (**C**) to previously identified biomarkers using bulk RNA sequencing. scRNA = single-cell RNA.

study identified a 4-gene biomarker of SSc correlating with severity of skin disease, IFI44, COMP, THBS1, and SIGLEC1 (1). Rice et al significantly expanded that signature, selecting genes based on microarray gene expression data to include IGFBP3, ADAM12, CTGF, IL13RA1, MS4A4A, THY1, WIF1, CD163, and CCL2 (2). In a subsequent study, Stifano et al identified OSMR, SERPINE1, and CD14 as prognostic biomarkers of progressive skin disease (3). We then examined the correlations of these bulk RNA sequencing-based correlates with the single-cell RNA sequencing-based biomarkers identified in the pan cell-type model (Figure 5A) and the keratinocyte- and fibroblast-specific models (Figures 5B and 5C, respectively). We visualized these correlations in a heatmap. We found that most of the previous biomarkers were significantly correlated with 1 or more of the biomarkers we identified from our analyses (Figures 5A–5C). However, this analysis additionally identified how these correlates

are distributed across 1 or more cell types. Furthermore, the correlates identified from the single-cell RNA-sequencing analyses are a minimal set that encapsulate the information from a wider range of relatively weaker proxies previously identified via bulk analyses.

We also examined the correlation of the newly identified cell type-specific biomarkers to the MRSS in previously generated bulk gene expression data. Indeed, most of these biomarkers identified in our cell type-specific model correlated with MRSS in bulk skin RNA. For example, STC2 and KRT6A correlated strongly with the MRSS, while IGFBP5 had a strong negative correlation. Others, like KRT5, correlated moderately (Supplementary Figure 2). The correlations of these genes in previous microarray studies are weaker than the correlations seen in individual cell populations. This highlights the value of studying gene expression by single-cell RNA sequencing, in which the

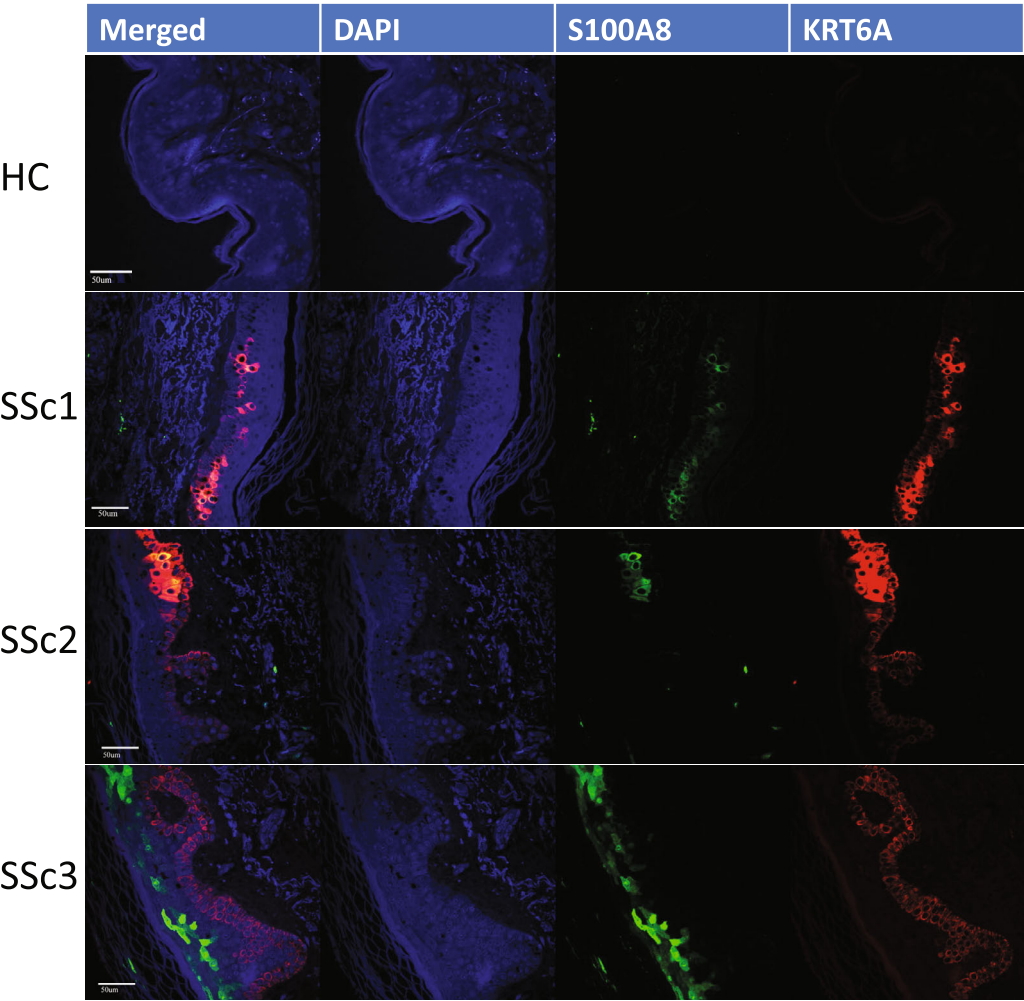


Figure 6. Staining of KRT6A and S100A8 in keratinocytes showing immunohistochemical detection of KRT6A and S100A8 in systemic sclerosis (SSc) skin compared to healthy control (HC) skin. KRT6A was detected mainly by keratinocytes in the basal layer, but also extending expression into suprabasal layers. S100A8 expression in most SSc skin largely overlapped with KRT6A staining. However, in 1 SSc skin sample, S100A8 was observed only in suprabasal keratinocytes. The intensity of staining of both proteins showed a positive trend correlating with the modified Rodnan skin score of $r = 0.75$ ($P = 0.086$) for KRT6A and $r = 0.74$ ($P = 0.093$) for S100A8.

RNA expression in individual cell types is not diluted by the expression in the other skin cell populations.

In order to understand better the processes marked by genes in the SFRP2-expressing fibroblast clusters, which we have shown are likely progenitors to myofibroblasts (9), we reclustered these cells and examined expression of key genes on feature plots. The expression of SAA1 in SFRP2+ cells marks a population of differentiating SSc fibroblasts (Supplementary Figure 1, <https://onlinelibrary.wiley.com/doi/10.1002/art.42536>). This population is co-correlated with expression of PRSS23, IGFBP2, and STC2 by SFRP2-expressing fibroblasts (Figure 3I). Each of these markers define slightly different SFRP2+ fibroblast populations that show transcriptomes that cluster increasingly close to myofibroblasts (Supplementary Figure 1).

Validation of the keratinocyte-specific biomarkers.

To examine the relationship between keratinocyte markers at the protein level, we stained SSc and healthy control skin for KRT6A and S100A8, 2 of the genes correlating with the MRSS in various keratinocyte populations. Staining for these markers was not detected in healthy control skin ($n = 3$), but was easily detected in 5 of 6 SSc skin samples (Figure 6). KRT6A was detected mainly by keratinocytes in the basal layer, but also extending expression into suprabasal layers (Figure 6). S100A8 expression in most SSc skin samples largely overlapped with KRT6A staining. However, in 1 SSc skin sample, S100A8 was observed only in suprabasal keratinocytes while KRT6A was expressed mainly in basal keratinocytes. The intensity of staining of both proteins showed a positive trend correlating with the MRSS ($r = 0.75$, $P = 0.086$ for KRT6A; $r = 0.74$, $P = 0.093$ for S100A8).

DISCUSSION

The clinical features of fibrotic, vascular, and immunologic disease have typically focused on investigation of fibroblasts, vascular, and immune cells in SSc. However, several recent studies have noted that the epidermis in SSc patients is hyperplastic, implicating keratinocytes in pathogenesis (29). Our current study shows that gene expression from multiple cell populations contributes to the most commonly used instrument for assessing clinical skin disease severity, the MRSS. While the biomarkers were identified at an individual time point, they provide insights into the cross-talk across multiple cell types involved in the pathogenesis process. Furthermore, beyond the individual biomarkers, the correlation networks provide more holistic insights into intracellular and extracellular coexpression networks underlying SSc pathogenesis.

In past studies, we described biomarkers correlating strongly with the MRSS that primarily represented genes from fibroblasts and macrophages (1,2). Surprisingly, gene expression by 3 genes expressed by suprabasal and proliferating

keratinocytes correlated more highly with the MRSS than genes from these populations. Previous studies have noted altered regulation of selected keratins at the protein level (30). KRT14 is normally restricted to the basal layer of keratinocytes, while KRT6 and KRT16 are not expressed at detectable levels in normal skin. In SSc, these show sharply up-regulated expression throughout the epidermis (KRT6 and KRT14) or suprabasal epidermis (KRT16). This dysregulated expression is reproduced in healing wounded skin (30). Similar changes have also been observed in radiation-induced skin fibrosis (31). In addition, in a previous microarray study of SSc skin, the “proliferative” signature associated with the “diffuse proliferative” patient subset largely originated from proliferating keratinocytes (32). This “proliferative” signature was shown to correlate with the MRSS. Our results strongly reinforce the relationship between altered keratinocyte differentiation and gene expression in SSc skin. Differential expression of KRT6A is associated with proliferating keratinocytes. Recent studies on epidermal equivalents show increased expression of KRT6 and KRT16 in 3-D cultures of epidermis generated from SSc patients compared to control skin (33). Although these investigators did not observe increased proliferation of basal cells, they attributed this to lower skin scores of the patients studied. KRT6 and KRT16 are also up-regulated in hyperproliferative psoriatic skin (34).

We recently identified PRSS23 as a marker for what appears to be a first step in myofibroblast differentiation (9). SAA1 marks a subgroup of PRSS23-expressing fibroblasts on uniform manifold approximation and projection (UMAP) analysis even more closely related to myofibroblasts. Although these biomarkers identify static cell populations, the co-coreregulated genes STC2, PRSS23, IGFBP2, and SAA1 appear to show progressive differentiation of SSc fibroblasts. This suggests that differentiation of SFRP2+ fibroblasts to myofibroblasts is a more gradual or multistep process than we had previously considered.

Down-regulated expression of IGFBP5 connected with down-regulated expression of WISP2. WISP2 has been observed previously to cluster with WIF1 in bulk microarray of SSc skin (32). WIF1 is strongly negatively associated with the MRSS in bulk microarray studies (2). WISP1 is expressed primarily by SFRP2+ fibroblasts. Down-regulated MYOC expression also correlates with IGFBP5, while CCL19, correlating positively with the MRSS, correlates negatively with IGFBP5. MYOC- and CCL19-expressing fibroblasts represent closely related fibroblast populations (9). MYOC is expressed in both adventitial and interstitial regions, whereas CCL19 is expressed mainly by adventitial fibroblasts. Thus, IGFBP5 appears to mark global changes in multiple fibroblast populations in SSc skin.

Our results indicate that altered keratinocyte gene expression is a strong biomarker of clinical disease severity

independent of previously defined biomarkers represented mainly by dermal fibroblasts. Applying LASSO to DEGs from single-cell populations, epidermal gene expression actually correlated more highly with the MRSS than dermal fibroblast markers. This observation potentially enables shave biopsies for monitoring disease activity in SSc patients. These correlations also open the possibility that the epidermis might be driving disease. Epidermal hyperplasia is seen in common skin diseases, such as atopic dermatitis and psoriasis. Comparison of SSc basal cell gene expression with these diseases may reveal similarities and differences in the underlying drivers of epidermal hyperplasia.

S100A9, the only myeloid hub gene in the combined-cell LASSO analysis, is an alarmin highly expressed by FCN+ and FCGR3A+ myeloid cells, which we have recently shown to be up-regulated in SSc skin (10). S100A9 co-correlates with metallothionein genes in fibroblasts via a strong correlation with THBS1. THBS1 is a strong biomarker for disease severity and 1 of 2 genes in our previously identified longitudinal biomarker of disease severity (2). Like S100A9, THBS1 is strongly up-regulated in all myeloid cell populations in SSc skin, but THBS1 is also strongly up-regulated in SFRP2-expressing fibroblasts, suggesting that a similar factor is activating both of these cell populations.

Myeloid cell expression of CCL18 correlates with pericyte expression of CXCL3. The prominent perivascular location of most myeloid cells in SSc skin suggests that these coregulated genes reflect an important cell–cell interaction. Both skin and serum levels of CCL18 decline rapidly and dramatically upon interleukin-6 (IL-6) inhibition (35). SAA1 expression by SFRP2-expressing fibroblasts is also coregulated with pericyte CXCL3. SAA1 is strongly induced by IL-6 (36), suggesting that IL-6 is contributing to the regulation of this hub of genes in several different cell types. Although we cannot find evidence in the literature indicating that CXCL3 is directly regulated by IL-6, lipopolysaccharides up-regulate pericyte expression of CXCL2, CXCL3, and IL-6 (37), and a role for Toll-like receptor 4 in SSc pathogenesis has been suggested in several studies (23,38). In addition, S100A8/S100A9 expressed by SSc skin myeloid cells stimulate keratinocyte secretion of CXCL2 and CXCL3, as well as IL-6 (39). Thus, this network suggests that innate immune stimulus, such as a Toll-like receptor 4 agonist or alarmin, activates IL-6, which in turn activates myeloid CCL18 and fibroblast SAA1 expression.

A limitation in the interpretation of the genes identified in this study is that it focuses only on the most highly, differentially expressed genes. These may not necessarily reflect the genes most actively driving pathogenesis. However, these genes do appear to provide important insights into cell–cell communications and robust markers of the disease process. Further investigation is necessary to translate these robust correlative relationships into actual mechanisms of disease pathophysiology. Additionally, we did not consider the relationships of the identified biomarkers to other clinical covariates for 2 reasons. First, the

coverage of clinical covariates is asymmetric and varies by cohort. Second, clinical covariates alone are not sufficient to predict disease severity. Nevertheless, we demonstrated that the identified cell type–specific biomarkers are sufficient to accurately predict SSc severity even when clinical covariates are not included in the analysis.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content. All authors approved the final version to be published. Mr. Berkowitz had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Lafyatis, Das.

Acquisition of data. Tabib, Sadej, Khanna, Fuschiotti, Lafyatis.

Analysis and interpretation of data. Berkowitz, Xiao, Das.

REFERENCES

- Farina G, Lafyatis D, Lemaire R, et al. A four-gene biomarker predicts skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheum* 2010;62:580–8.
- Rice LM, Ziemek J, Stratton EA, et al. A longitudinal biomarker for the extent of skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheumatol* 2015;67:3004–15.
- Stifano G, Sornasse T, Rice LM, et al. Skin gene expression is prognostic for the trajectory of skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheumatol* 2018;70:912–9.
- Pendergrass SA, Lemaire R, Francis IP, et al. Intrinsic gene expression subsets of diffuse cutaneous systemic sclerosis are stable in serial skin biopsies. *J Invest Dermatol* 2012;132:1363–73.
- Whitfield ML, Finlay DR, Murray JI, et al. Systemic and cell type-specific gene expression patterns in scleroderma skin. *Proc Natl Acad Sci USA* 2003;100:12319–24.
- Xu X, Ramanujam M, Visvanathan S, et al. Transcriptional insights into pathogenesis of cutaneous systemic sclerosis using pathway driven meta-analysis assisted by machine learning methods. *PLoS One* 2020;15:e0242863.
- Kerick M, González-Serna D, Carnero-Montoro E, et al. Expression quantitative trait locus analysis in systemic sclerosis identifies new candidate genes associated with multiple aspects of disease pathology. *Arthritis Rheumatol* 2021;73:1288–300.
- Apostolidis SA, Stifano G, Tabib T, et al. Single cell RNA sequencing identifies HSPG2 and APLNR as markers of endothelial cell injury in systemic sclerosis skin. *Front Immunol* 2018;9:2191.
- Tabib T, Huang M, Morse N, et al. Myofibroblast transcriptome indicates SFRP2^{hi} fibroblast progenitors in systemic sclerosis skin. *Nat Commun* 2021;12:4384.
- Xue D, Tabib T, Morse C, et al. Expansion of Fcγ receptor IIIa–positive macrophages, ficolin 1–positive monocyte-derived dendritic cells, and plasmacytoid dendritic cells associated with severe skin disease in systemic sclerosis. *Arthritis Rheumatol* 2022;74:329–41.
- Das J, Devadhasan A, Linde C, et al. Mining for humoral correlates of HIV control and latent reservoir size. *PLoS Pathog* 2020;16:e1008868.
- Goetghebuer T, Smolen KK, Adler C, et al. Initiation of antiretroviral therapy before pregnancy reduces the risk of infection-related hospitalization in human immunodeficiency virus–exposed uninfected infants born in a high-income country. *Clin Infect Dis* 2019;68:1193–203.

13. Ackerman ME, Das J, Pittala S, et al. Route of immunization defines multiple mechanisms of vaccine-mediated protection against SIV. *Nat Med* 2018;24:1590–8.
14. Sadanand S, Das J, Chung AW, et al. Temporal variation in HIV-specific IgG subclass antibodies during acute infection differentiates spontaneous controllers from chronic progressors. *AIDS* 2018;32:443–50.
15. Das J, Fallon JK, Yu TC, et al. Delayed fractional dosing with RTS,S/AS01 improves humoral immunity to malaria via a balance of polyfunctional NANP6- and Pf16-specific antibodies. *Med* 2021;2:1269–86.
16. Suscovich TJ, Fallon JK, Das J, et al. Mapping functional humoral correlates of protection against malaria challenge following RTS,S/AS01 vaccination. *Sci Transl Med* 2020;12:eabb4757.
17. Lu LL, Das J, Grace PS, et al. Antibody Fc glycosylation discriminates between latent and active tuberculosis. *J Infect Dis* 2020;222:2093–102.
18. Assabban A, Dubois-Vedrenne I, Van Maele L, et al. Tristetraprolin expression by keratinocytes protects against skin carcinogenesis. *JCI Insight* 2021;6:e140669.
19. Dayem MA, Moreilhon C, Turchi L, et al. Early gene expression in wounded human keratinocytes revealed by DNA microarray analysis. *Comp Funct Genomics* 2003;4:47–55.
20. Pan SC, Li CY, Kuo CY, et al. The p53-S100A2 positive feedback loop negatively regulates epithelialization in cutaneous wound healing. *Sci Rep* 2018;8:5458.
21. Xu Z, Cheng C, Kong R, et al. S100A8 and S100A9, both transcriptionally regulated by PU.1, promote epithelial-mesenchymal transformation (EMT) and invasive growth of dermal keratinocytes during scar formation post burn. *Aging (Albany NY)* 2021;13:15523–37.
22. Nazari B, Rice LM, Stifano G, et al. Altered dermal fibroblasts in systemic sclerosis display podoplanin and CD90. *Am J Pathol* 2016;186:2650–64.
23. Bhattacharyya S, Wang W, Morales-Nebreda L, et al. Tenascin-C drives persistence of organ fibrosis. *Nat Commun* 2016;7:11703.
24. Dai H, Wang L, Li L, et al. Metallothionein 1: a new spotlight on inflammatory diseases [review]. *Front Immunol* 2021;12:739918.
25. Park Y, Zhang J, Cai L. Reappraisal of metallothionein: clinical implications for patients with diabetes mellitus [review]. *J Diabetes* 2018;10:213–31.
26. Dubé A, Harrisson JF, Saint-Gelais G, et al. Hypoxia acts through multiple signaling pathways to induce metallothionein transactivation by the metal-responsive transcription factor-1 (MTF-1). *Biochem Cell Biol* 2011;89:562–77.
27. Aiba S, Tabata N, Ohtani H, et al. CD34+ spindle-shaped cells selectively disappear from the skin lesion of scleroderma. *Arch Dermatol* 1994;130:593–7.
28. Wu M, Pedroza M, Lafyatis R, et al. Identification of cadherin 11 as a mediator of dermal fibrosis and possible role in systemic sclerosis. *Arthritis Rheumatol* 2014;66:1010–21.
29. Nikitorowicz-Buniak J, Shiwen X, Denton CP, et al. Abnormally differentiating keratinocytes in the epidermis of systemic sclerosis patients show enhanced secretion of CCN2 and S100A9. *J Invest Dermatol* 2014;134:2693–702.
30. Aden N, Shiwen X, Aden D, et al. Proteomic analysis of scleroderma lesional skin reveals activated wound healing phenotype of epidermal cell layer [review]. *Rheumatology (Oxford)* 2008;47:1754–60.
31. Song J, Zhang H, Wang Z, et al. The role of FABP5 in radiation-induced human skin fibrosis. *Radiat Res* 2018;189:177–86.
32. Milano A, Pendergrass SA, Sargent JL, et al. Molecular subsets in the gene expression signatures of scleroderma skin [published correction appears in *PLoS One* 2008;3(10)]. *PLoS One* 2008;3:e2696.
33. Russo B, Borowczyk J, Boehncke WH, et al. Dysfunctional keratinocytes increase dermal inflammation in systemic sclerosis: results from studies using tissue-engineered scleroderma epidermis. *Arthritis Rheumatol* 2021;73:1311–7.
34. Thewes M, Stadler R, Korge B, et al. Normal psoriatic epidermis expression of hyperproliferation-associated keratins. *Arch Dermatol Res* 1991;283:465–71.
35. Khanna D, Denton CP, Jähreis A, et al. Safety and efficacy of subcutaneous tocilizumab in adults with systemic sclerosis (faSScinate): a phase 2, randomised, controlled trial [published correction appears in *Lancet* 2018;391:1356]. *Lancet* 2016;387:2630–40.
36. Hagihara K, Nishikawa T, Isobe T, et al. IL-6 plays a critical role in the synergistic induction of human serum amyloid A (SAA) gene when stimulated with proinflammatory cytokines as analyzed with an SAA isoform real-time quantitative RT-PCR assay system. *Biochem Biophys Res Commun* 2004;314:363–9.
37. Guijarro-Muñoz I, Compte M, Álvarez-Cienfuegos A, et al. Lipopolysaccharide activates Toll-like receptor 4 (TLR4)-mediated NF- κ B signaling pathway and proinflammatory response in human pericytes. *J Biol Chem* 2014;289:2457–68.
38. Bhattacharyya S, Kelley K, Melichian DS, et al. Toll-like receptor 4 signaling augments transforming growth factor- β responses: a novel mechanism for maintaining and amplifying fibrosis in scleroderma. *Am J Pathol* 2013;182:192–205.
39. Nukui T, Ehama R, Sakaguchi M, et al. S100A8/A9, a key mediator for positive feedback growth stimulation of normal human keratinocytes. *J Cell Biochem* 2008;104:453–64.